CSCI 384

Computational Linguistics

Fall 2015 MWF 3:15–4:20 SCI 184

http://cs.wheaton.edu/~tvandrun/cs384

Thomas VanDrunen ☎630-752-5692 630-639-2255 ✉Thomas.VanDrunen@wheaton.edu
Office: SCI 163 Office hours: MTuWThF 9:15-10:15 am; Th 1:30-3:30 pm.

Contents

CATALOG DESCRIPTION. An exploration of big ideas in computational linguistics, natural language processing, and/or language technologies. Language models, n-grams, information theory and entropy, and semantics. Applications of computational linguistics such as part-of-speech tagging, authorship attribution, automatic translation, and sentiment analysis. Prerequisite: CSCI 345 (non-majors without the prerequisite may enroll with departmental approval).

TEXTBOOK. Daniel Jurafsky and James Martin, *Speech and Language Processing,* second edition, Prentice Hall, 2009.

(Also recommended:) Steven Bird et al, *Natural Language Processing with Python,* O'Reilly, 2009. (A copy will be available in the computer science student lounge and also at http://nltk.org/book/)

OBJECTIVES. At the end of this course students should

- demonstrate understanding of the principle definitions, goals, methods, contributions, and applications of the field of computational linguistics (and/or natural language processing, language technologies, statistical language modeling. . .);
- demonstrate a basic competence in some of the tools and methods of the field (the NLTK library, for example);
- demonstrate understanding of the algorithms and mathematics powering the language technologies in the world around us; and
- describe and assess how language technologies are affecting popular culture, business and marketing, scholarship in the humanities (including biblical scholarship), our understanding of human language, and the ministry of the gospel.

Additionally (and informally) students will have learned something about programming/algorithms, linguistics, and statistics (hopefully all three).

This course will support students' development towards the computer science major's learning outcomes # 6 (Practical—tools) and # 7 (Applied—integrative issues). It further will reinforce and assess the retention of things related to #4 (Formal—algorithmic analysis) and # 5 (Practical—programming).

OUTLINE.

The following is a *conceptual* outline of the course to explain the relationships among the various topics and themes. It is not a *temporal* outline; background, core, and applied topics will be pursued in parallel. For the sequence of coverage and schedule, see the course website.

I. Background
   - A. Linguistic
     - i. Phonetic
     - ii. Lexical
     - iii. Syntactic
     - iv. Semantic
     - v. Rhetorical
   - B. Statistical
     - i. Basic probability
     - ii. Bayes's theorem
     - iii. Distributions
     - iv. Hidden Markov models
   - C. Computational
     - i. Regular expressions
     - ii. Finite-state automata
     - iii. Dynamic programming
     - iv. Machine learning

II. Core
   - A. Words
     - i. Types, tokens, n-grams
     - ii. Heads and tails, hapax legomena, function words, stop words
     - iii. Rank vs frequency; Zipf's law
     - iv. Lexical semantics
   - B. Language models
     - i. Maximum likelihood, relative frequency
     - ii. Perplexity
     - iii. Smoothing
   - C. Information theory
     - i. Goals and basic theory
     - ii. Entropy
     - iii. Linguistic insights from entropy

III. Applications
   - A. Standard
     - i. POS tagging
     - ii. Spelling correction
     - iii. Parsing
   - B. Analytical
     - i. Sentiment analysis
     - ii. Stylometry and authorship attribution
     - iii. Information retrieval
   - C. Synthetic
     - i. Machine translation
     - ii. Text generation
     - iii. Discourse agents

Course procedures

How we do this course. For each class period, students will have a portion of the text book and/or supplemental material to read ahead of time (this may be divided into "read carefully" and "skim" portions). Most of classtime will be in a lecture format, with occasional group work on problems (including coding problems). Most of the outside work in this course will go into the projects—four well-defined "regular" projects plus one open-ended term project. The midterm and final exam will assess students' mastery and retention of the formal side of the topic stream.

There may also be quizzes and short assignments to enforce and exercise the concepts in class and to give a feel for what the exam will be like. (These would be rolled into the "Participation" portion of the grading scheme below.)

Applied topics. The last two or three weeks of the semester are reserved for applied topics that will be selected with input from the students. Possible topics include authorship attribution and other applications of stylometry, language usage trends, sentiment analysis, discourse analysis, question answering, information retrieval, text generation, discourse agents, machine translation, and language learning tools.

Projects. Expected topics and seasons for the projects: Project 1 (regular expressions and NLTK), early September; project 2 (edit distance), late September, early October; project 3 (Hidden Markov Models), mid/late October; project 4 (parsing), early November. Subject to change. Mid-November through the end of the semester is reserved for work on the term project.

Term project. Students will choose a term project suited to their interest either exploring an applied topic or extending an earlier project, approved by the instructor. The specific form of the project will vary, but at minimum it will include some amount of research into related work, the creation of relevant software, the acquiring of a dataset, and a report. The project's initial proposal should be no later than Oct 21, with refinement and final approval by Oct 30. Students may choose to work in pairs (or triples) with the understanding that their project would be appropriately more ambitious.

Grading. The final exam block is Thursday, Dec 17, 10:30–12:30 am.

instrument	*weight*
Participation	15
Regular projects (4)	35 (total)
Midterm	15
Term project	20
Final exam	15

"Participation" includes short assignments, readings, quizzes, lab activities, contribution to class discussion, and anything else that may come up.

Policies etc

Academic Integrity. Collaboration among students enrolled in the course is permitted on all projects and assignments. Obtaining solutions to projects through electronic or other media is strictly prohibited. Any ideas obtained through electronic, print, or other media must be cited as they would in a research paper. Any violations will be handled though the college's disciplinary process.

Late assignments. Students are allowed two late days for close-ended projects (either one project two days late or two projects each one day late). Beyond that, late projects will not normally be accepted. The term project will have a hard deadline (unless otherwise noted, that will be the last day of classes) and will not be accepted late.

Attendance. Students are expected to attend all class periods. It is courtesy to inform the instructor when a class must be missed.

EXAMINATIONS. Students are expected to take all tests, quizzes, and exams as scheduled. In the case where a test must be missed because of legitimate travel or other activities, a student should notify the instructor no later than one week ahead of time and request an alternate time to take the test. In the case of illness or other emergency preventing a student from taking a test as scheduled, the student should notify the instructor as soon as possible, and the instructor will make a reasonable accommodation for the student. The instructor is under no obligation to give any credit to students for tests to which they fail to show up without prior arrangement or notification in non-emergency situations. The final exam is Thursday, Dec 17, 10:30 AM. I do not allow students to take finals early (which is also the college's policy), so make appropriate travel arrangements.

SPECIAL NEEDS. *Institutional statement:* Wheaton College is committed to providing reasonable accommodations for students with disabilities. Any student with a documented disability needing academic adjustments is requested to contact the Academic and Disability Services Office as early in the semester as possible. Please call 630.752.5941 or send an e-mail to `jennifer.nicodem@wheaton.edu` for further information.

My own statement: Whenever possible, classroom activities and testing procedures will be adjusted to respond to requests for accommodation by students with disabilities who have documented their situation with the registrar and who have arranged to have the documentation forwarded to the course instructor. Computer Science students who need special adjustments made to computer hardware or software in order to facilitate their participation must also document their needs with the registrar in advance before any accommodation will be attempted.

GENDER-INCLUSIVE LANGUAGE. The college requires the following statement to be included on all syllabi: *For academic discourse, spoken and written, the faculty expects students to use gender inclusive language for human beings.*

OFFICE HOURS. I try to keep a balance: Stop by anytime, but prefer my scheduled office hours. Any time my door is closed, it means I'm doing something uninterruptible, such as making an important phone call. Do not bother knocking; instead, come back in a few minutes or send me an email. Be aware that this semester I teach a class meeting immediately after ours (the 3:15-4:20 slot), so I will not be available during what is often a popular office-hour time.

DRESS AND DEPORTMENT. Please dress in a way that shows you take class seriously—more like a job than a slumber party. (If you need to wear athletic clothes because of activities before or after class, that's ok, but try to make yourself as professional-looking as possible.) If you must eat during class (for schedule or health reasons), please let the instructor know ahead of time; we will talk about how to minimize the distraction.

ELECTRONIC DEVICES. I do not allow the use of laptops or tablets for note-taking or textbook reference except in special, pre-arranged circumstances. (Only once have I ever allowed a student to use a tablet in class for referencing a textbook. I have never allowed a student to use a laptop or tablet in class for note-taking.) Trying out programming concepts on your own during class time (for example) is not productive because it takes you away from class discussion. You cannot multi-task as well as you think you can. Please make sure all electronic devices are silenced and put away. ***Text in class and DIE.***